DOI:
10.1016/S2468-1253(16)30007-3

Document Version
Peer reviewed version

THELANCETGASTROHEP-D-16-00002R2

DEVELOPMENT AND VALIDATION OF A NOVEL DYNAMIC OUTCOME PREDICTION MODEL FOR PARACETAMOL-INDUCED ACUTE LIVER FAILURE; A COHORT STUDY

Brief Title: Outcome Prediction Model for Paracetamol ALF

William Bernal MD[+]*, Yanzhong Wang PhD*[A], James Maggs MBBS, Christopher Willars FRCA, Elizabeth Sizer FRCA, Georg Auzinger MD, Nicholas Murphy FRCA[B], Damian Harding FRCAP[B], Ahmed Elsharkawy PhD[B], Kenneth Simpson MD[C], Fin Stolze Larsen MD[D], Nigel Heaton FRCS [F], John O'Grady MD [F], Roger Williams MD [E, F] and Julia Wendon MD [F]

***Joint First Authors**

Institute of Liver Studies, Kings College Hospital, London UK

[A] Division of Health and Social Care Research, King's College London,

[B] Departments of Liver Medicine, Anaesthesia and Critical Care, Queen Elizabeth Hospital, Birmingham UK

[C] Department of Hepatology, Edinburgh Royal Infirmary, Edinburgh UK

[D] Department of Hepatology, Rigshospitalet University Hospital Copenhagen, Denmark.

[E]Foundation for Liver Research, London

[F] Full Professors

[+]Corresponding Author

Dr William Bernal

Liver Intensive Therapy Unit, Institute of Liver Studies, Kings College Hospital, London SE5 9RS, UK

Telephone: 0203 299 4458

Email: william.bernal@kcl.ac.uk

12/05/2016

Research in Context

Evidence before this study

We searched PubMed and Medline with the terms “acute liver failure” and “fulminant hepatic failure” to identify studies and publications in addition to those familiar to the authors or cited by guidelines. We used no date restrictions. Reference lists in identified reports were also reviewed to further identify studies of relevance to this study. This included identification of 3 meta-analyses of prognostic criteria in acute liver failure (ALF), but the majority of studies reporting survival outcomes were case series from single centres with cohorts of limited size and open to reference bias. Available evidence suggests that survival of patients with paracetamol-induced ALF managed with medical care alone has progressively increased over time, and that existing criteria used to select candidates for emergency liver transplantation (ELT) may no longer be effective.

Added value of this Study

This study presents novel criteria for the assessment of prognosis in paracetamol-induced ALF, derived and validated in a very large multi-centre patient cohort. The models developed reflect the current outcomes of the illness and utilise readily available clinical variables and standardised definitions, with sequential assessment to assess changing prognosis over time. For the first time the models enable individualised and updated survival predictions to be made and address key practical issues in relation to clinical use.

Implications of all the available evidence

ELT may be life-saving in selected patients with paracetamol-induced ALF, however many of those transplanted using existing criteria may have survived with medical management alone. The use ELT for this indication requires comprehensive re-evaluation.

Abstract

Background

Early, accurate prediction of survival is central to management of patients with paracetamol-induced acute liver failure (P-ALF) to identify those needing emergency liver transplantation (ELT). Current prognostic tools are confounded by recent improvements in outcome independent of ELT, and constrained by static binary outcome prediction. We developed a simple prognostic tool that reflected current outcomes, generating a dynamic updated estimation of risk of death.

Methods

1028 Patients with P-ALF managed at intensive care units in the United Kingdom (London, Birmingham, Edinburgh) and Denmark (Copenhagen) were studied. Prognostic models were developed, excluding transplanted patients, using Cox Proportional-Hazards in a derivation (n=350), and tested in initial (n=150) and recent external (n=412) validation datasets. Mortality was estimated in those patients who had ELT (n=116) Model discrimination was assessed using area under receiver operating characteristic curve (AUROC) and calibration by Root-Mean-Square Error (RMSE). Admission (Day 1 (D1)) variables age, Glasgow Coma Scale, arterial pH and lactate, creatinine, INR, and cardiovascular failure were used to derive an initial predictive model, with a second (D2) model including additional changes in INR and lactate.

Findings

Applied to the derivation set the AUROC for 30-day survival using the D1 model was 0·92 (95% confidence interval 0·88-0·96) and D2 0·93 (0·88-0·97). In the initial validation set AUROC was 0·89 (0·84-0·95) and 0·90 (0·85-0·95) with RMSE 0·1642 and 0·0626 and in the external validation set, 0·91 (0·87-0·94) and 0·91 (0·88-0·95) and 0·079 and 0·107. Applied to 116 patients who underwent ELT, median predicted 30-day survival was 51% (33-85).

Interpretation

We present the development and validation of novel high performance statistical models to support decision-making in patients with P-ALF. The models show very good discrimination and calibration, confirmed in independent datasets and suggest that many of those transplanted using existing criteria may have survived with medical management alone. The role and indications for ELT in P-ALF requires re-evaluation.

Funding: Unrestricted Grant from Liver Research Foundation

Background

Acute liver failure (ALF) is a rare critical illness, and in many countries, including the United Kingdom, paracetamol-induced hepatotoxicity its most frequent cause [1,2]. The illness may follow a rapidly progressive course with severe hepatic necrosis quickly followed by the development of encephalopathy, multiple extra-hepatic organ failures and death [3]. However, the condition is one in which there is marked variation in clinical course. In some patients' recovery with medical therapy alone may be possible despite severe multiple organ failure, whilst in others survival may be impossible without emergency liver transplantation (ELT) [4,5]. The early and accurate evaluation of expected prognosis is key to effective management, to enable successful transplantation in the narrow window of opportunity for those who will benefit from it, and avoiding unnecessary surgery in those who will survive with medical therapy alone. Tools available for prognostic evaluation are available, but their limitations are increasingly apparent [6-8].

Prognosis in paracetamol-induced ALF (P-ALF) is most commonly assessed using the Kings College Criteria (KCC). The criteria were derived from the analysis of patients managed in a single centre between 1973 and 1987 and have been in use to select transplant candidates for more than two decades[6]. Poor prognosis criteria include an arterial pH of <7.3 after volume resuscitation, or the combined findings of high grade encephalopathy, creatinine of more than 300 µMol/l and INR ≥6.5 occurring within a narrow time frame [6]. Early meta analysis of the KCCs diagnostic performance in P-ALF suggested high specificity but more limited sensitivity; the introduction of arterial blood lactate concentration as a supplemental marker was proposed to address this issue [7-9]. However, the increasing success of non-transplantation medical therapies alone is likely to have affected the performance of the KCC in patients with P-ALF as in the last three decades marked improvements in survival with medical care alone have occurred for many aetiologies of ALF, particularly those cases resulting from paracetamol-induced hepatotoxity[3]. These improvements in non-transplant outcomes have not been reflected in changes in recognised indications for ELT.

Experience has also shown practical issues in the use of the KCC. They were intended for use in a transplant centre and not early after presentation in the emergency room where intravenous fluid resuscitation has not been undertaken and the effect of high blood levels of paracetamol may contribute to a reversible lactic acidosis [10,11]. They perform less accurately in accidental or intentional staggered overdoses of the drug than after single time point deliberate ingestions, and be difficult to interpret as their component variables may be confounded by alteration by medical interventions [12]. As they provide a binary outcome prediction rather than a continuum of risk, their application may also be difficult, and they do not address a key clinical dilemma in the waitlisted

patient who shows signs of improvement – is it better to remove them from the list or to proceed with transplantation?

In exploring the development of a novel model for the prediction of death without transplantation in patients with paracetamol-induced ALF, our study design sought to address the limitations of the present KCC. By studying more recent cohorts we hoped to reflect the current outcomes of the illness and develop a model utilising readily available clinical variables and standardised definitions, with sequential assessment to assess changing prognosis over time [13]. The objective was to develop a decision support tool that was simple to use and gave a continuous and updated estimation of risk of death rather than a static binary outcome prediction. We hoped to avoid the reference bias that may complicate studies of prognosis in ALF by excluding those patients who underwent transplantation, and to assess reproducibility and transportability by the use of multiple validation cohorts.

Patients and Methods

Patients with severe paracetamol-induced hepatotoxicity managed at specialist liver transplantation intensive care units (ICU) in the United Kingdom and Denmark were studied. In all cases a history of drug ingestion and / or detectable blood paracetamol was present with the exclusion of other causes of acute liver injury. Other inclusion criteria included an INR of ≥ 1·5 and absence of a previous history and clinical / radiologic findings of liver disease. The primary derivation and initial internal validation test cohorts were derived from consecutive non-transplanted patients with severe paracetamol-induced hepatotoxicity admitted over the period 2000-2012 to the Liver Intensive Therapy Unit (LITU) at Kings College Hospital (KCH), London UK.

External validation cohorts included 151 non-transplanted patients over the period 2011-13 from the Rikshopitalet liver unit in Copenhagen, Denmark, 90 patients from 2008-14 admitted to the Scottish Liver Transplant Unit at the Edinburgh Royal Infirmary, 72 patients from 2004-13 treated at the ICU of Queen Elizabeth Hospital, Birmingham UK and a further 99 patients treated at KCH LITU between 2012 and 2014. Features on admission of these cohorts are shown in table 2. Validating mortality in paracetamol-induced ALF is seldom an issue as almost without exception non-transplanted death results during a single hospital admission, most commonly during the first ICU week , with rapid recovery seen in those who survive [14].

A common approach to clinical management was applied in all units, with ELT considered in patients who fulfilled the standard KCC. Standard medical care applied has been detailed elsewhere [3]. In brief, patients developing encephalopathy grade ≥3 were intubated, sedated and mechanically

ventilated. Guided restoration of circulating volume was commenced immediately on admission and utilized invasive hemodynamic monitoring. Coagulopathy was not supported unless active haemorrhage was present [15]. Norepinephrine was the primary vasopressor used and dobutamine the primary ionotropic agent with adjunctive use of intravenous low dose hydrocortisone and vasopressin. Renal replacement therapy (RRT) utilised continuous veno-venous hemofiltration (CVVHF). Indications for its use including not only those standard for patients with acute kidney injury with anuria but also for relative oliguria, metabolic stabilization and control of acidosis and hyper-ammonaemia. Sedation utilized fentanyl and propofol infusions with rare use of paralysis. Treatment for intra-cranial pressure crises was with bolus intravenous mannitol, hypertonic saline and increased sedation using thiopentone in refractory cases. Intravenous N-acetyl Cysteine (NAC) was administered to all patients with an infusion of 100 mg/kg/24 hrs for a maximum of 5 days or until the INR was <2.

Datasets and Statistical Methods

The derivation and initial validation patient datasets were drawn from the LITU clinical database where demographic, physiologic and laboratory parameters are prospectively collected daily on all patients by specialist audit nurses. This included those required for the Sequential Organ Failure Assessment (SOFA) score and Acute Physiology and Chronic Health Evaluation (APACHE) scores (web appendix page 1). Fewer than 1% of cases had missing values.

Standardised scores were utilised assessing level of consciousness and grade of encephalopathy using the Glasgow Coma Scale (GCS), and cardiovascular dysfunction by the SOFA cardiovascular component score (SOFA CVS; web appendix page 2), with a score of 3 or above taken to represent cardiovascular failure (SOFA CVS Failure) [3,16,17]. GCS was assessed in patients who had not received sedative agents, and if used these was taken as the lowest score prior to administration, except if signs of intracranial hypertension were present [3,18]. Survival time was from date of admission to date of death. Continuous variables are summarised as mean (standard deviation) or median (Interquartile range (IQR)) and categorical data as count (percentage). Student's t-test, Mann-Whitney U and Wilcoxon signed-rank tests were used to test differences in continuous variables where appropriate, and the Chi-square test used for proportions. Multiple survival analyses were undertaken using Cox Proportional-Hazards models to determine the prognostic and predictive value of demographic factors and clinical variables. All-cause mortality was the primary event studied and transplanted patients were excluded from model development and primary testing. Variable selection and model fitting were conducted through backward stepwise regression based on p-values. The model starts with a range of clinical variables of important prognostic value

suggested by literature review, and then goes through extensive backward stepwise model/variable selection process. The final models consisted of variables most strongly associated with death in the multiple Cox proportional hazards regressions. The proportional-hazards assumption for each covariate was tested using the scaled Schoenfeld residuals. The proportionality test showed that all the covariates followed the proportional-hazards assumption (*P*>0.05).

To develop and validate the predictive survival models, we divided the KCH dataset into derivation set (n=350) and initial validation set (n=150) with random case selection. Survival models using day 1 and day 2 were first built from derivation dataset separately and then validated in the validation dataset. The performance and predictive accuracy of the models were assessed using receiver operating characteristic (ROC) analysis. The ROC curves and values were compared among different models and also between derivation and validation stages. Model calibration was assessed comparing observed and predicted survival using Root-Mean-Square Error (RMSE) [19]. All tests were 2-tailed, and p-value < 0·05 was considered statistically significant. ROC values with 95% CI for all the survival models and hazard ratios (HR) with 95% confidence Intervals (CI) for the chosen clinical variables were calculated. Statistical analyses were performed with statistical software R, version 2·11·1 and SPSS version 22·0·0. All data was fully deidentifed before exchange and analysis and its use was approved by the Research Ethics Committee of Kings College Hospital.

Results

Characteristics of the derivation and initial validation sets are shown in table 1.

Derivation Set

Seventy-eight (22%) of 350 patients in the derivation set died during hospitalisation; median day of death was 9 (IQR 2-14) days after admission. Non-survivors were older and on admission had evidence of more severe liver dysfunction with higher INR and arterial lactate levels and worse extra-hepatic organ failure with more severe encephalopathy, cardiovascular and renal dysfunction (table 3). Of note, 49 of 106 (46%) patients with GCS ≤9 on D1 died as compared to 29 of 244 (12%) with GCS above this threshold (Relative Risk 3·9 (2·6-5·8), p<0.0001).

After extensive model fitting and variable selection, admission (D1) variables of age, GCS, arterial pH and lactate, creatinine, INR, and SOFA CVS failure were identified as best clinical predictors. Hazard ratios for these component variables on admission are shown in table 4 (a) and the predictive equation in supplemental materials 2. Based on D1 model, we explored a day 2 (D2) model with additional changes in all clinical variables between day 1 and day 2, and after backward stepwise variable selection found only changes in blood lactate and INR to be significantly associated with

survival. A dynamic (D2) survival model was thus further developed based on clinical variables on day 1 plus changes in blood lactate level and INR between day 1 and 2 to reflect the changing patterns in these critical variables (table 4 (b), web appendix page 3). Exploration of models using data up to 7 days after admission did not demonstrate significant improvements in model performance above the D2 model.

When applied to the initial validation set, AUROC for prediction of death at 7, 15 and 30 days using the D1 model was 0·95 (95% Confidence interval 0·91-0·99), 0·94 (0·90-0·97) and 0·92 (0·88-0·96) respectively, and for the D2 model 0·96 (0·93-1·0), 0·95 (0·91-0·98) and 0·93 (0·88-0·97) (figure 1 (a) and (b)). Assessment of calibration utilising RMSE in prediction of 30-day survival gave values of 0·1123 and 0·1317 for the D1 and D2 models respectively.

Initial Validation Set

Thirty-five (23%) of the 150 patients in the initial validation set died during hospitalisation. Using the D1 model AUROC for the prediction of death at 7, 15 and 30 days was 0·93 (0·86-1·00), 0·91 (0·85-0·96) and 0·89 (0·84-0·95), and with the D2 model 0·94 (0·88-1·0), 0·91 (0·86-0·96) and 0·90 (0·85-0·95) (figure 2 (a), (b)). Assessment of calibration utilising RMSE in prediction of 30-day survival gave values of 0·1642 and 0·0626 for the D1 and D2 models respectively (figure 2, (c), (d)). Based on the proposed survival models, individual survival curves were calculated and updated during the first 2 days of ICU admission (figure 3).

External Validation Sets

Characteristics of the four validation sets differed significantly from one another and from the primary cohorts. Median age of the Danish patients was higher and the Birmingham cohort had more severe acidosis and lactate and creatinine, more severe encephalopathy and higher mortality (table 2). Overall, D1 single component variables were missing in 19 of 412 cases (5%) and D2 variables in 141 (34%). Where D2 INR and lactate variables were missing, values were categorised as showing no improvement.

AUROC and RMSE for the individual validation sets are shown in table 5. In the combined external validation sets, AUROC for 30-day survival for the D1 and D2 models was 0·91 (0·87-0·94) and 0·91 (0·88-0·95) and RMSE 0·079 and 0·107 respectively (Figure 4).

Patients Transplanted

During the study periods of the derivation and validation sets, a total of 116 patients underwent ELT; 84 at KCH, 23 at Birmingham and 8 in Edinburgh. Comparison of admission variables of those

patients who died with those transplanted showed significant differences, with the former being older and with higher arterial lactate levels, and lower GCS and INR values (web appendix page 4). Using the D1 model, median predicted 30-day survival for the transplanted cohort was 51% (95% CI 33-85) and where severe HE (GCS≤9) was present (n=70) predicted survival was 36% (18-73).

Discussion

In this report, we present the development and validation of high performance statistical models to support decision making in the care of patients with P-ALF. The models proposed show very good discrimination and calibration, confirmed on application to multiple independent external validation datasets with patients with a range of illness severity, and show considerable promise as prognostic tools. We developed models both on (D1) and after (D2) admission as their combined use provides practical support for key clinical decisions not addressed by existing single time point prognostic criteria. In management of P-ALF the first two days are critically important for the selection of transplant candidates; early prediction is required and is directly clinically relevant [14]. Experience has shown that the clinical condition of patients with P-ALF may change rapidly after admission and initial therapy – with improvement in some and deterioration in others. The two stage assessment of prognosis permits quantitation of this change, objectively reassessing prognosis in those patients who show signs of improvement. Assessment of test performance showed both models to have high discrimination but importantly calibration was improved in the D2 model. This sequential approach, derivation from recent patient cohorts and the utilisation of novel prognostic important variables represent clear advances beyond the original KCC.

There has been criticism within the statistical literature on the use of stepwise regression for model fitting relating to its biases and suboptimal results. However, we used stepwise regression for its simplicity and in conjunction with expert opinion to decide which variables to include in the model. Our predictive models were robustly trained and validated on both internal and external datasets, which resulted in excellent model performance. Nevertheless, there are alternatives to stepwise regression which include Partial Least Squares Regression (PLS) and Least Absolute Shrinkage and Selection Operator (LASSO). These alternatives can overcome some of the shortcomings of stepwise regression although they have their own limitations [20]. We will explore these novel methods in future research to further enhance the performance of our predictive models.

We chose to exclude those patients who underwent ELT from model derivation and initial validation. Valid criticism can be made both of the inclusion or exclusion of transplanted patients in this process and both have been utilised in the literature. Our principal concern was of risk of reference bias given the dramatic improvements in survival with medical management alone in patients with P-ALF,

and the possibility that detail of transplantation practice varied between centres [3,5] .Whilst this approach could introduce changes in cohort composition, it is unlikely to have yielded a less 'sick' sample as in key respects those who later died were more severely ill. Interpretation of the predictions of a survival model derived in a non-transplanted population applied to one that was transplanted must be with caution. However this survival estimate is derived from the analysis of a population of patients with illness of the same cause, which cases that had a range of disease severity that will have included patients of similar acuity and with a model that utilises variables derived solely from the pre-transplant phase of illness.

When applying the D1 model to the patients transplanted during the study period, the median predicted survival was just more than 50% but fell to 36% when only those with severe encephalopathy (HE) were considered. These observations reinforce the key prognostic importance of its development and that transplantation should not be considered in the absence of high grade HE. Importantly they also suggest that an appreciable proportion of those transplanted might have survived with medical management alone. These findings quantify the marked improvements in survival independent of liver transplantation that have occurred over time and suggest that its place as an intervention in P-ALF requires comprehensive re-evaluation [21 3].

In considering the application of these models in a clinical setting, limitations to their potential practical use must be assessed. In other critical-illness scoring systems, use for individual patient outcome prediction has historically been with caution, as the primary purpose of most such scores was for group outcome assessment and quality assurance [22-24]. However, there is clear precedent within Hepatology for the use of scores, including MELD, for individual decision making in relation to transplantation – and in the widely accepted use of the KCC and other poor prognosis criteria for selection of patients with ALF for emergency liver transplantation [24,25].

We see these models as tools to quantify the risk of death and support expert clinical judgement and decision making; experience in other areas of acute and critical care medicine suggests that combining an objective prognosis measure with a physicians clinical estimate results in the most accurate assessment of actual prognosis [23,26,27]. Further, the sequential assessment our models provide is likely to be of benefit as in critically ill patients with and without liver disease, trends in illness severity provide additional prognostic value over single static determinations [23,28-30]. Rather than considering a single time point survival estimate, transplantation wait listing decisions might best be made from observations of the dynamic course of the illness. An obvious issue is that of the threshold of estimated survival that should trigger wait listing; a figure of 25% demonstrated in recent reports of early transplantation for acute alcoholic hepatitis provides a useful parallel [31]. A

website has been developed to use the new paracetamol prediction models in a form that is accessible, easy to use and whose output will provide real-time estimates of expected survival, whilst accumulating a further prospective confirmatory validation set.

In developing these models we deliberately chose not to place any reliance upon the reported timing of drug ingestion in our patient selection and survival modelling. In practice, this information is often inconsistent or unavailable and in the case of overdoses staggered over days defining a specific time point of ingestion impossible. Even without reliance upon this information, the model functioned well. However, it is important to recognise that these patients were assessed in liver transplantation centres usually days after drug ingestion, with established and significant hepatic necrosis and after receiving initial resuscitative measures at their receiving hospitals prior to transfer. Use of the models to predict survival in patients soon after overdose or at early after first presentation has not been assessed.

These models were primarily designed to identify those patients whose survival would be enhanced by ELT. However, identification as having a poor prognosis does not necessarily mean that survival will be improved by liver transplantation. The models are designed to predict survival without transplantation but not survival after surgery, where factors not considered in our model are of prognostic importance [32]. In many cases contraindications to ELT may be present and here the models may rather serve to guide patients and family members in the expected outcome of illness.

Author Contributions: the Study was conceived by WB and JW. YW and WB developed the predictive model and WB wrote the first draft of the paper. WB, JM, NM, AE, DH, KS and FSL contributed data and all authors contributed in detail to the writing of the final version of the manuscript.

Role of the Funding Source: this study was supported by an unrestricted grant from the Foundation for Liver Research. The Funding source had no role in study design, collection, analysis or interpretation of the data or in the writing of the report. Authors had access to the raw data from their individual centres and WB and YW to all de-identified datasets. The corresponding author had final responsibility to submit for publication.

Declaration of Interest: The authors declared no conflicts of interest

.

Table 1. Demographics, clinical and laboratory findings and outcome of Derivation and Initial Valdidation Cohorts.

		Derivation		**Validation**	
n		350		150	
Age (years)		37 (27-48)		36 (26-47)	
Gender (F)		188 (54%)		94 (63%)	
Died (%)		78 (22%)		35 (23%)	
Day 1					
HE Grade ≥3		149 (43%)		64 (43%)	
Glasgow Coma Score		13 (8-15)		14 (8-15)	
CVS Failure (n (%))*		83 (24%)		35 (23%)	
Mean Arterial Pressure (mmHg)		74 (63-95)		70 (61-90)	
INR		4·0 (2·6-6·2)		3·8 (2·8-6·0)	
Bilirubin (µMol/l)		68 (44-97)		72 (44-104)	
AST (IU/l)		5776 (2601-9809)		5215 (2089-9315)	
Creatinine ((µMol/l)		162 (92-267)		172 (93-281)	
Arterial pH		7·40 (7·31-7·4)		7·40 (7·31-7·43)	
Arterial Lactate (mMol/l)		2·8 (1.9-4.5)		2·6 (1·8-4·9)	
Day 2					
INR		3.0 (2.1-4.6)		3.0 (2.2-4.9)	
Arterial Lactate (mMol/l)		2.0 (1.4-3.2)		2.1 (1.4-3.2)	

Note; Data is median (interquartile range) or n (%). INR; International Normalised Ratio, AST; Aspartate Transaminase, HE; Hepatic Encephalopathy. * Sequential Organ failure Assessment (SOFA) Cardiovascular score ≥3

Table 3. Comparison of demographics and admission characteristics of surviving and non-surviving patients of the derivation dataset.

		Died		**Survived**	
n		78		272	
Age (years)		46 (37-53)		35 (25-45)	
Gender (F)		43 (55%)		145 (53%)	
INR		5.3 (3.3-8.7)		3.7 (2.6-5.7)	
Bilirubin (µMol/l)		60 (40-97)		69 (46-97)	
AST (IU/l)		6539 (3221-9999)		5545 (2454-9498)	
Creatinine ((µMol/l)		204 (151-289)		141 (82-248)	
Arterial pH		7·26 (7·17-7·37)		7·40 (7·36-7·42)	
Arterial Lactate (mMol/l)		7·6 (4·2-12·8)		2·4 (1·6-3·5)	
HE Grade ≥3		59 (76%)		90 (33%)	
Glasgow Coma Score		7 (6-11)		15 (10-15)	
CVS Failure (n (%)*		51 (65%)		32 (12%)	
Mean Arterial Pressure		63 (54-71)		78 (66-99)	

Note; Data is median (interquartile range) or n (%). INR; International Normalised Ratio, AST; Aspartate Transaminase, HE; Hepatic Encephalopathy. * Sequential Organ failure Assessment (SOFA) Cardiovascular score ≥3

Table 2. Demographics, admission (day 1) laboratory findings and outcome of external validation sets.

	Copenhagen	Edinburgh	Birmingham	Kings	All
n	151	90	72	99	412
Age (years)	52 (36-61)	35 (27-46)	40 (31-46)	39 (30-50)	42 (31-53)
Gender (F)	93 (61%)	49 (54%)	39 (54%)	59 (59%)	240 (58%)
INR	3·0 (2·2-4·3)	5·3 (3·9-7·2)	4·9 (3-7·1)	6·2 (3·8-9·7)	3·9 (2·7-6·0)
Arterial pH	7·42 (7·36-7·46)	7·41 (7·31-7·47)	7·29 (7·19-7·38)	7·40 (7·35-7·45)	7·40 (7·30-7·45)
Creatinine (µMol/l)	82 (56-153)	119 (73-216)	186 (96-297)	146 (76-273)	118 (66-236)
Arterial Lactate (mMol/l)	2·6 (1·7-4·9)	2·9 (1·8-5·2)	5·6 (3·9-10·9)	3·3 (2·3-7·3)	3·5 (2·1-7·1)
Glasgow Coma Score	15 (8-15)	15 (14-15)	9 (9-10)	14 (9-15)	14 (9-15)
CVS Failure*	26 (17%)	17 (19%)	57 (79%)	59 (60%)	159 (39%)
Died (%)	23 (15%)	13 (14%)	36 (50%)	27 (27%)	99 (24%)

Note; Data is median (interquartile range) or n (%). INR; International Normalised Ratio. * Sequential Organ failure Assessment (SOFA) Cardiovascular score ≥3

Table 4 (a). Clinical Predictors included in admission (D1) predictive model in the derivation dataset (n=350).

		Hazard Ratio	95% CI		p
Age (5yrs)		1·07	(0·97-1·18)		0·168
Day 1					
CVS Failure*		2·41	(1·39-4·17)		0·002
Glasgow Coma Score		0·90	(0·84-0·98)		0·009
Arterial pH		0·06	(0·01-0·61)		0·018
Log (Creatinine (per 10 units))		1·74	(1·12-2·7)		0·013
Log (INR)		1·53	(1·04—2·24)		0·029
Sqrt (Arterial Lactate)		2·01	(1·53-2·63)		<0·001

Table 4 (b). Clinical Predictors included in Dynamic (D2) predictive model in the derivation dataset (n=350).

		Hazard Ratio	95% CI		p
Age (5yrs)		1·07	(0·98-1·18)		0·146
Day 1					
CVS Failure		3·14	(1·8-5·49)		<0·001
Glasgow Coma Score		0·90	(0·83-0·97)		0·005
Arterial pH		0·09	(0·01-0·95)		0·045
Log (Creatinine(per 10 units))		1·70	(1·08-2·67)		0·022
Log (INR)		1·97	(1·33—2·91)		<0·001
Sqrt (Arterial Lactate)		2·01	(1·53-2·63)		<0·001
Day 2					
Lower Arterial Lactate		0·31	(0·13-0·69)		0·0045
Lower INR		0·54	(0·29-0·99)		0·046

Note * Sequential Organ failure Assessment (SOFA) Cardiovascular score ≥3. Sqrt; Square root.

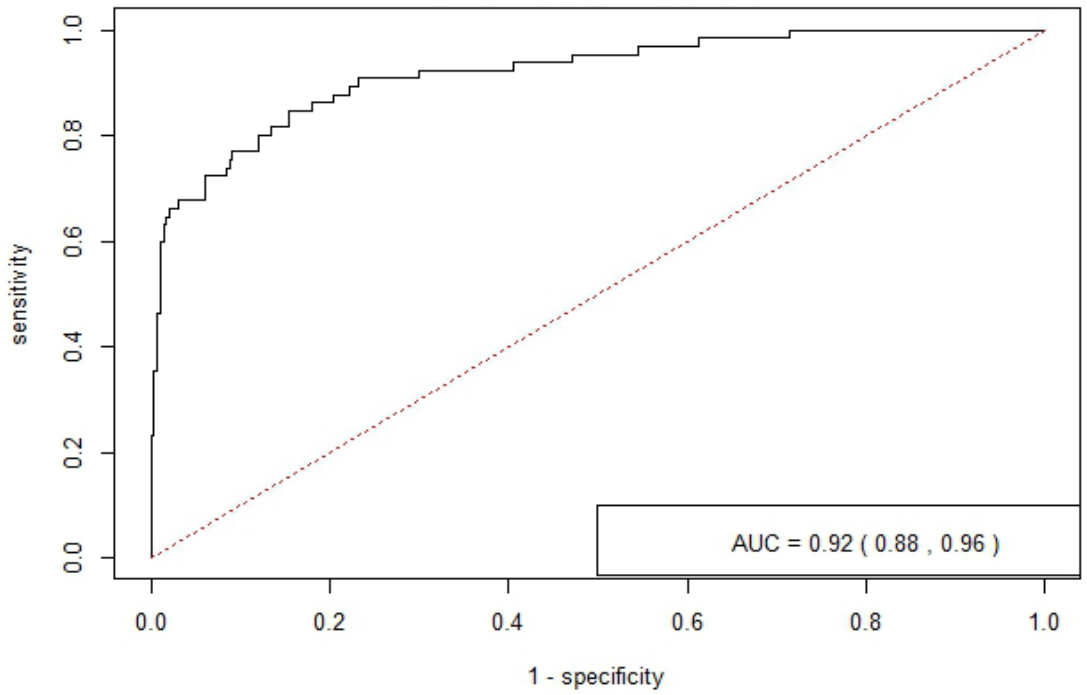
Table 5. Diagnostic test performance of D1 and D2 models for predicting 30-day survival applied to individual and combined external validation sets.

		D1 Model				D2 Model		
Set	n	AUROC	(95% CI)	RMSE		AUROC	95% CI	RMSE
Copenhagen	151	0·93	(0·88-0·98)	0·140		0·94	(0·87-1·00)	0·208
Edinburgh	90	0·88	(0·79-0·96)	0·281		0·89	(0·81-0·97)	0·239
Birmingham	72	0·84	(0·74-0·94)	0·053		0·83	(0·73-0·93)	0·111
Kings	99	0·92	(0·85-0·99)	0·117		0·93	(0·86-0·99)	0·185
Combined	412	0·91	(0·87-0·94)	0·079		0·91	(0·88-0·95)	0·107

Note: AUROC; Area Under Receiver Operating Characteristic Curve, RMSE; Root-Mean-Square Error.

Figure 1. Area under receiver operating characteristic curve (AUROC) for Kings College Hospital (KCH) Datasets.

(a) D1 Model KCH Derivation Set (n=350).



(b) D2 Model KCH Derivation Set (n=350).

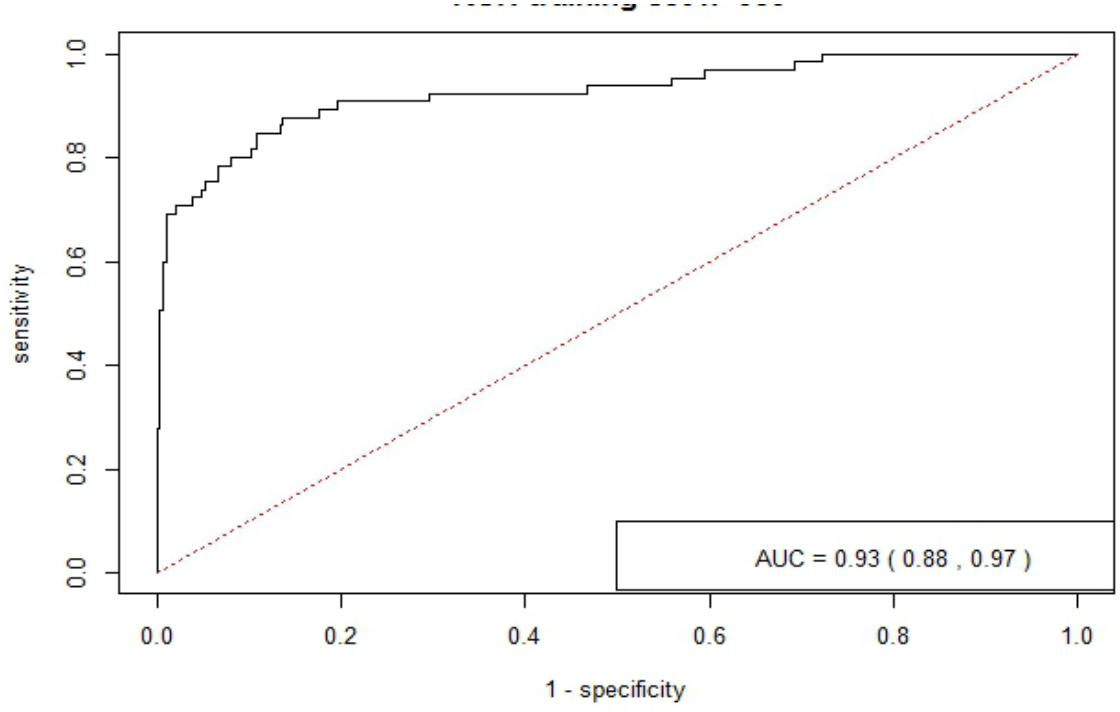
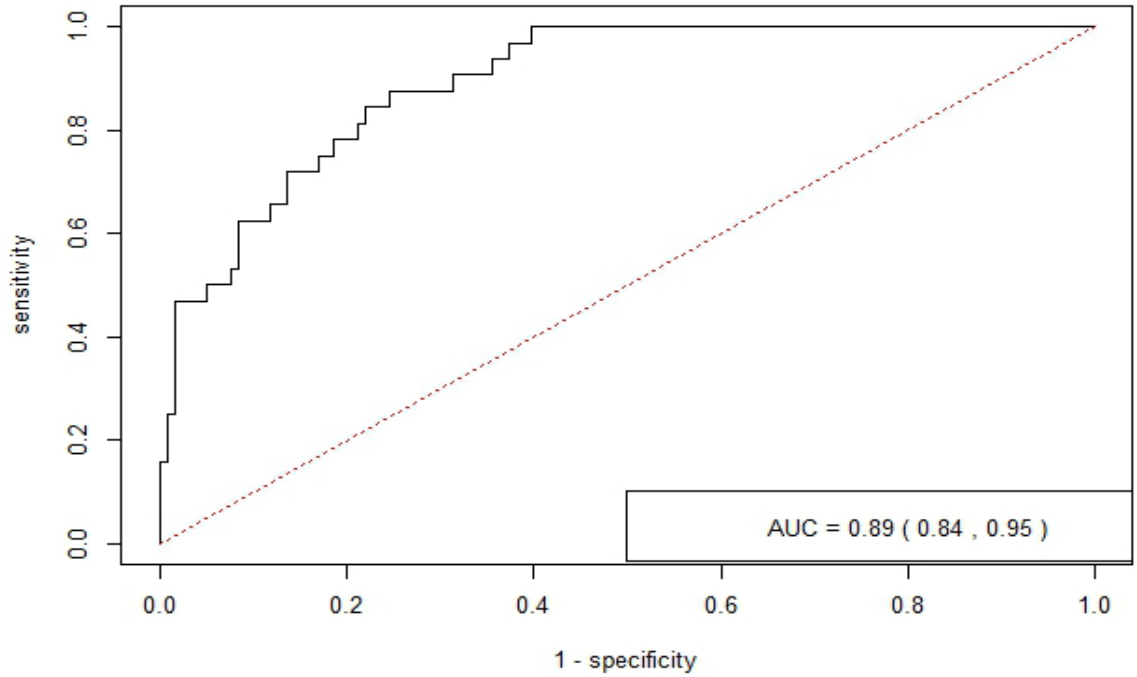
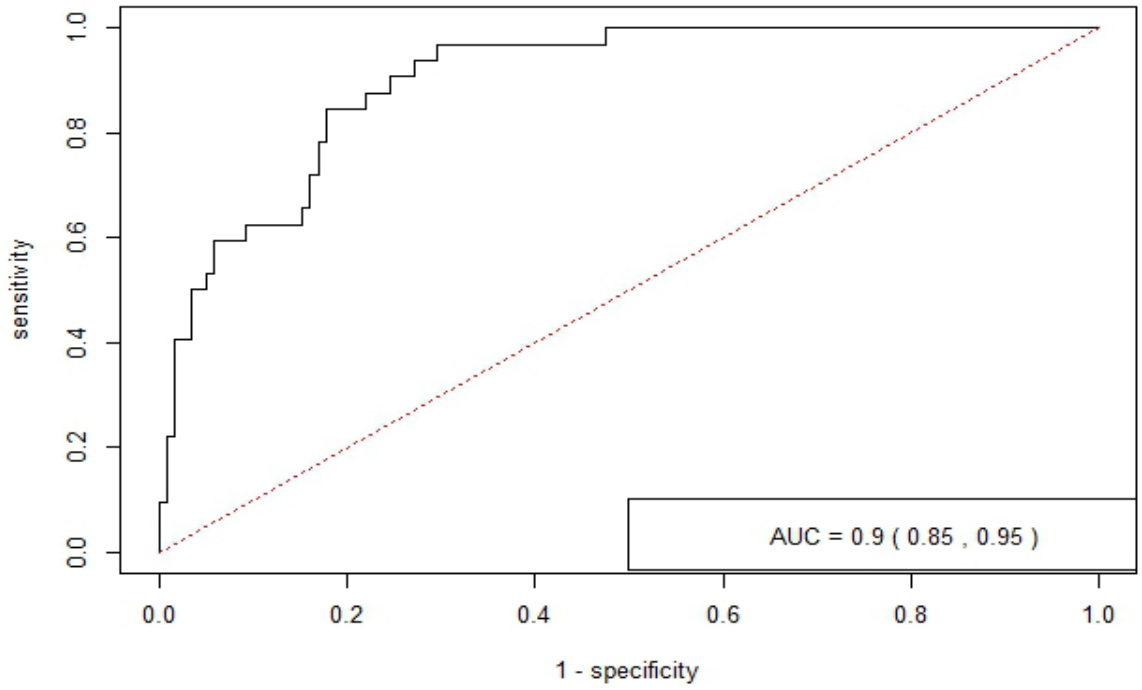


Figure 2. Discrimination and Calibration of D1 and D2 models in KCH Validation set.

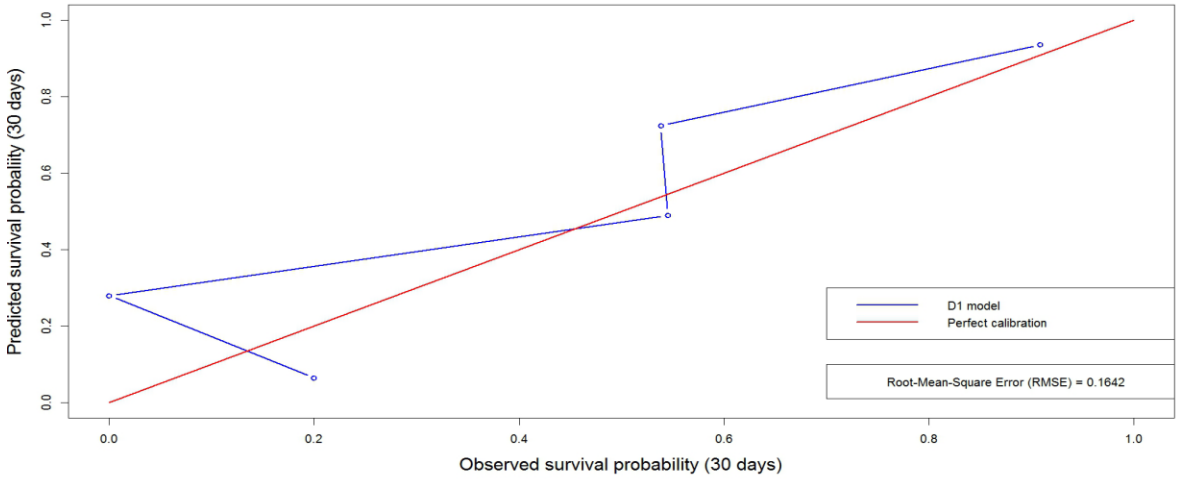
(a) AUROC for D1 Model (n=150).



(b) AUROC for D2 Model (n=150).



(c) Calibration curve for D1 Model (n=150)



(d) Calibration curve for D2 Model (n=150)

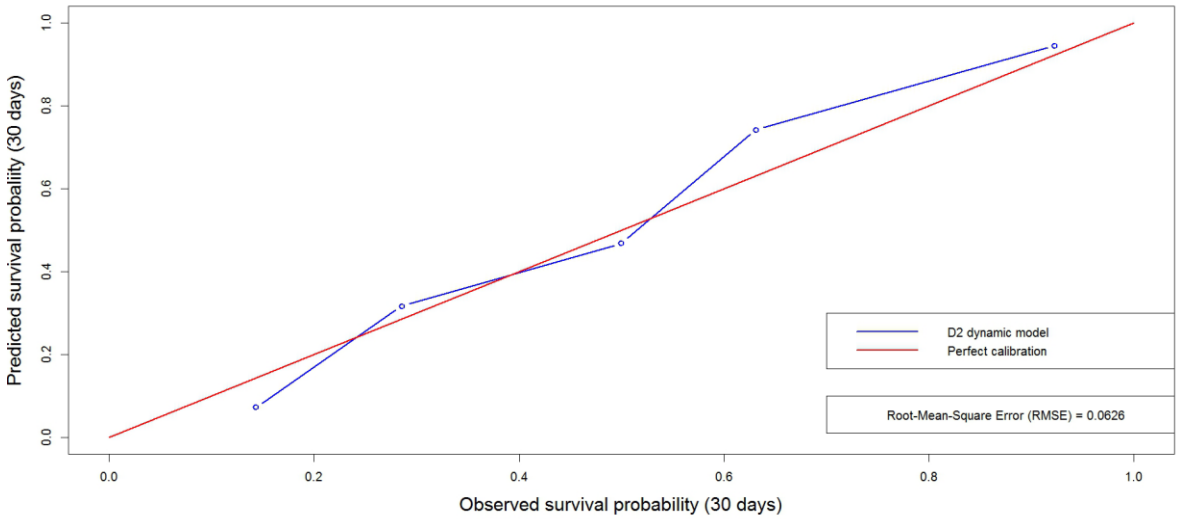
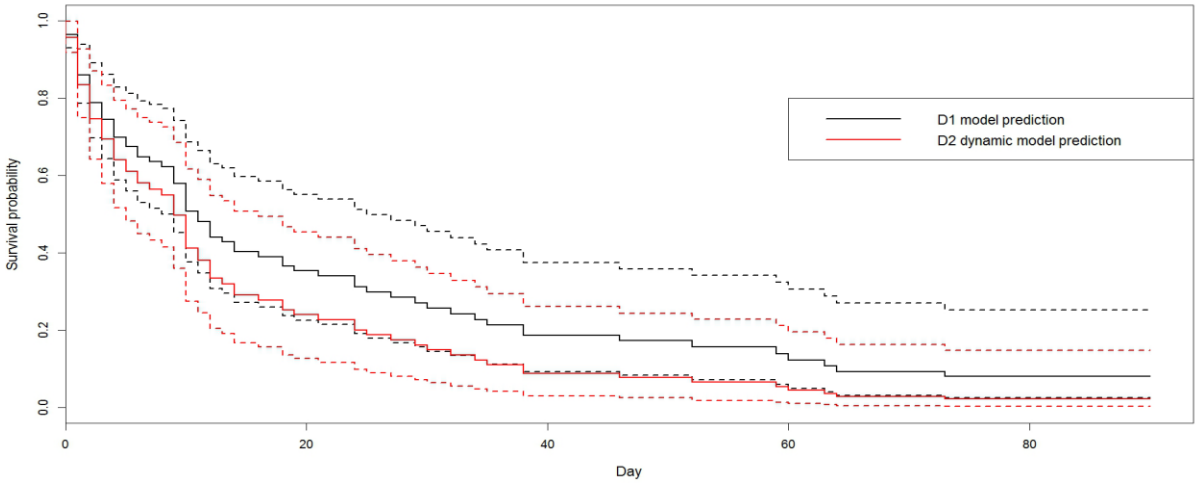
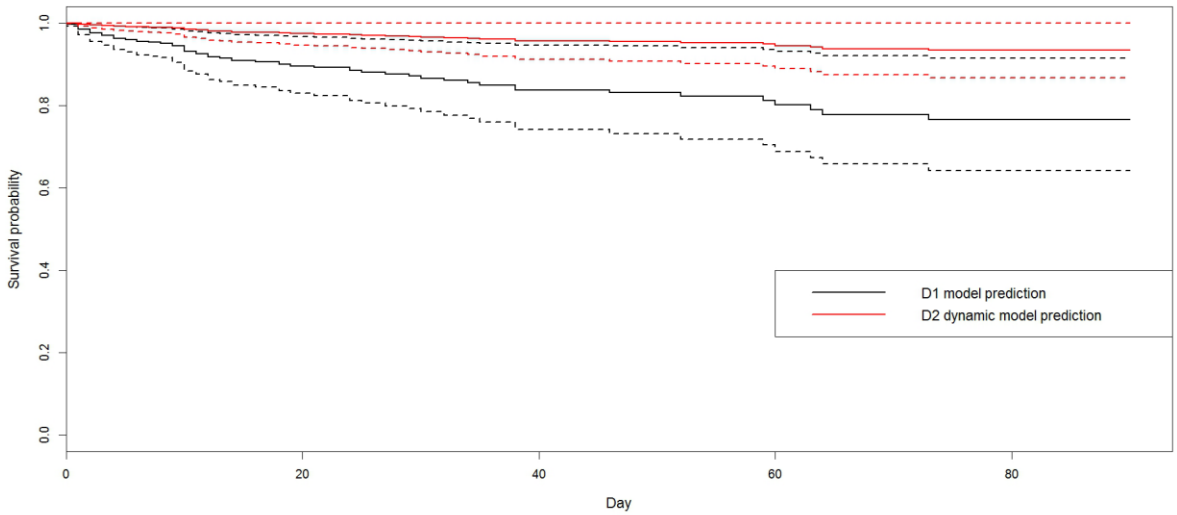


Figure 3. Predicted Survival curves by D1 and D2 models for individual patients from KCH derivation set.

(a) Patient died at day 23.



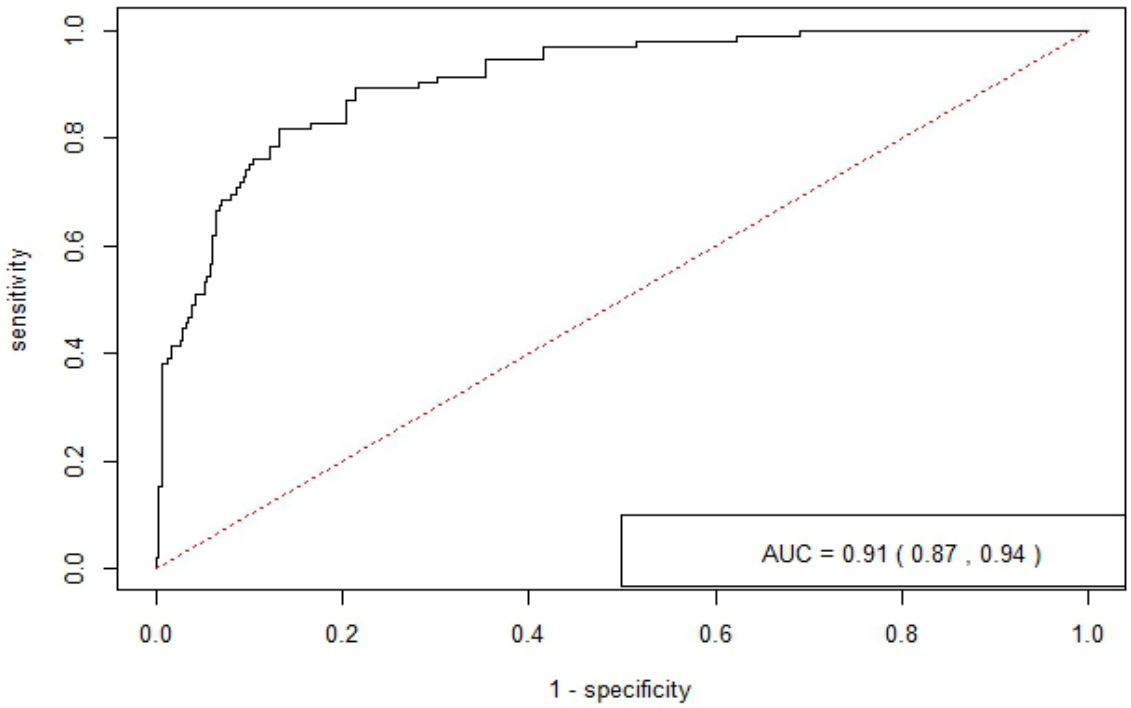
(b) Patient survived to hospital discharge.



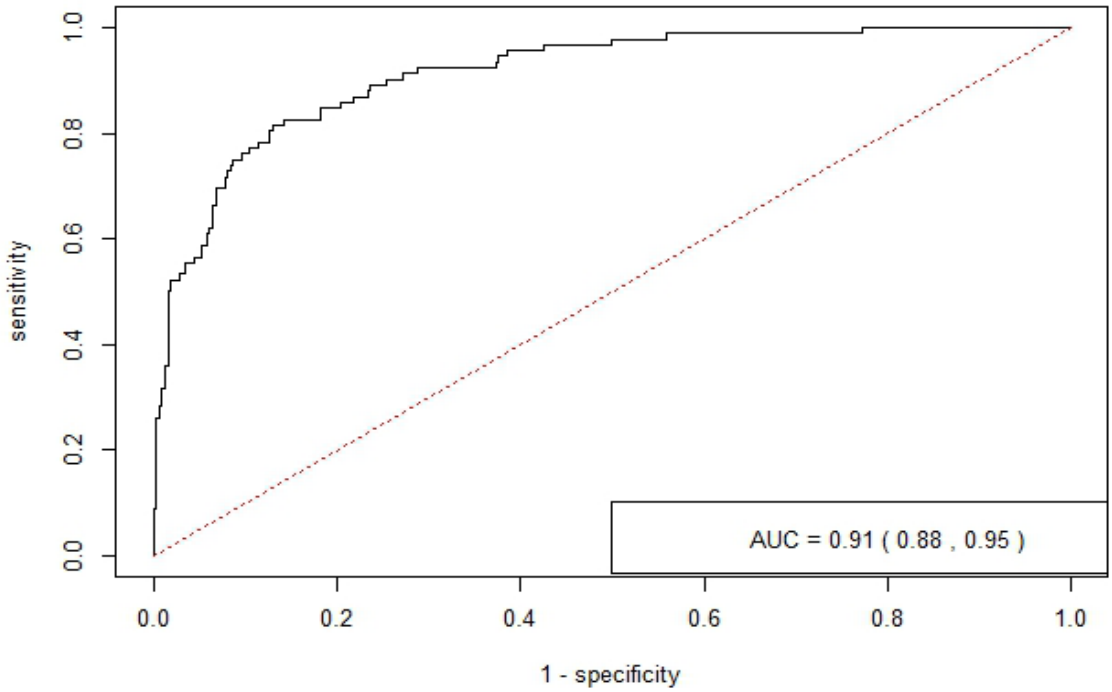
Note; Hashed lines represent 95% confidence intervals for survival estimate.

Figure 4. Discrimination and Calibration of D1 and D2 models in combined external validation sets (n=412).

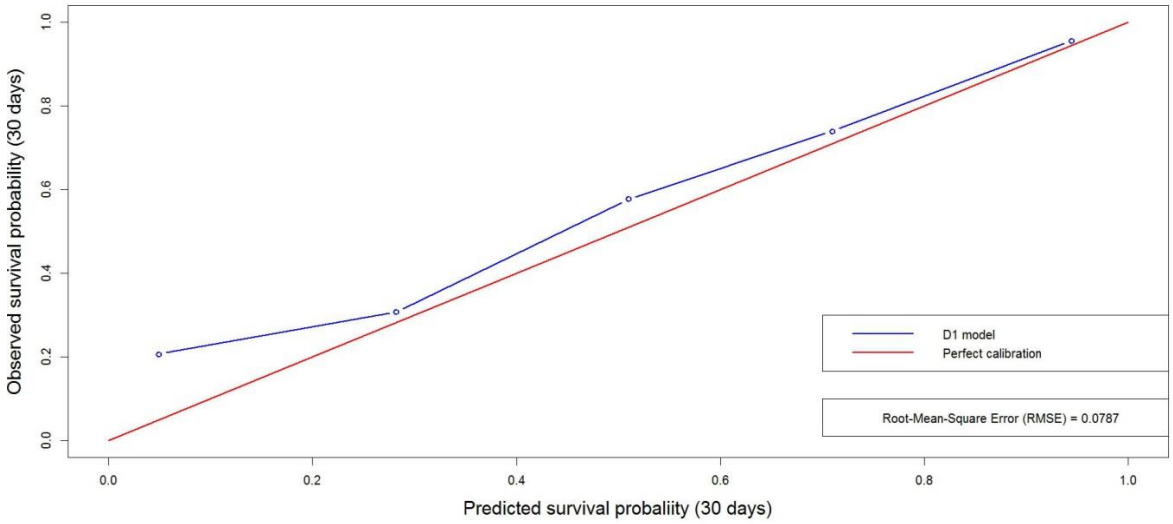
(a) Area under receiver operating characteristic curve (AUROC) for D1 Model



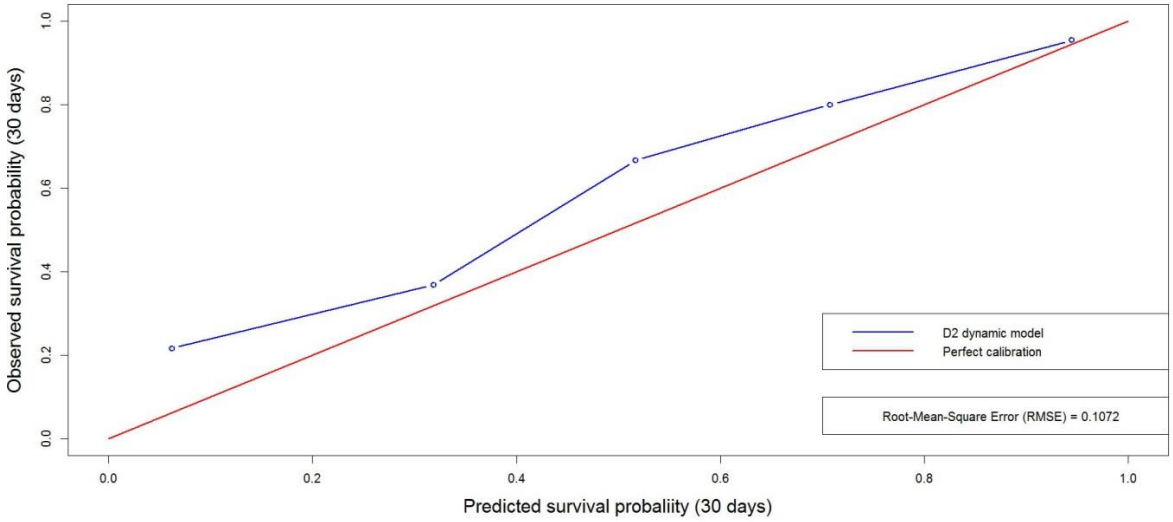
(b) Area under receiver operating characteristic curve (AUROC) for D2 Model


sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.91 (0.88 , 0.95)

(c) Calibration Curve for D1 model



(c) Calibration Curve for D2 model



Supplemental Materials 1.

Variables assessed in predictive model derivation:

Age

Gender

Heart rate

Mean Arterial Pressure

Systolic Blood Pressure

Diastolic Blood Pressure

SOFA Cardiovascular Failure

Respiratory rate

PaO_2/FiO_2 ratio

$PaCO_2$

Arterial pH

Arterial [HCO_3]

Arterial [lactate]

Blood concentrations:

Sodium

Potassium

Creatinine

Urea

Leucocyte count

Platelet count

INR

Bilirubin

Aspartate Transaminase

Albumin

Glucose

Glasgow Coma Scale

Urine Output

Supplemental Materials 2.

Predictive Equations for D1 and D2 Models

Predictive equation for D1 model:

Survival function (integration of hazard function over time 0 to t) is used to calculate the survival probability *S(t)* for any case at time t:

Based on the estimation results (Hazard ratios) from the training dataset, we were able to calculate the survival probability for any patient at time t:

$$S(t) = S_0(t)^K$$

$$\mathrm{K} = \mathrm{e}^{\mathrm{PI}} = \exp(b_{age} * age + b_{cardio} * Cardio + b_{GCS} * GCS + b_{pH} * pH + b_{creatin} * \log(creatin) + b_{INR} * \log(INR) + b_{lactate} * \sqrt[2]{lactate}) = 1.07^{age} + 2.41^{cardio} + 0.90^{GCS} + 0.06^{pH} + 1.74^{\log(creatin)} + 1.53^{\log(INR)} + 2.01^{\sqrt[2]{lactate}}$$

$S_0(t)$ can be estimated through the baseline cumulative hazard *$H_0(t)$* and need help of statistical software.

Predictive equation for D2 model:

Based on the estimation results (Hazard ratios) from the training dataset, we were able to calculate the survival probability for any patient at time t:

$$S(t) = S_0(t)^K$$

$$\mathrm{K} = \mathrm{e}^{\mathrm{PI}} = \exp(b_{age} * age + b_{cardio} * Cardio + b_{GCS} * GCS + b_{pH} * pH + b_{creatin} * \log(creatin) + b_{INR} * \log(INR) + b_{lactate} * \sqrt[2]{lactate} + b_{dynamic\ lactate} * (dynamic\ lactate) + b_{dynamic\ INR} * (dynamic\ INR)) = 1.07^{age} + 3.14^{cardio} + 0.90^{GCS} + 0.09^{pH} + 1.70^{\log(creatin)} + 1.97^{\log(INR)} + 2.01^{\sqrt[2]{lactate}} + 0.31^{(dyamic\ lactate)} + 0.54^{(dynamic\ INR)}$$

$S_0(t)$ can be estimated through the baseline cumulative hazard $H_0(t)$ and need help of statistical software.

Supplementary materials 3.

Comparison of demographics and admission laboratory findings of patients who died or underwent transplantation.

		Died		Transplanted	p
n		185		116	
Age (years)		44 (35-54)		33 (24-40)	<0·001
Gender (F)		104 (56%)		75 (65%)	0·15
INR		5·0 (3·3-8·7)		6·9 (4·4-10)	0·003
Arterial pH		7·27 (7·15-7·38)		7·28 (7·20-7·38)	0·39
Glasgow Coma Score		8 (6-10)		9 (7-12)	0·005
Creatinine (µMol/l)		200 (128-282)		197 (129-298)	0·96
Arterial Lactate (mMol/l)		8·1 (4·3-12·7)		5·9 (4-9·7)	0·01
CVS Failure		151 (82%)		89 (77%)	0·30

Note; INR; International Normalised Ratio. Patients transplanted are 16%, 24% and 8% of the patient samples from Kings, Birmingham and Edinburgh cohorts respectively.

Supplementary Materials 4.

The Sequential Organ failure Assessment (SOFA) Cardiovascular score

				SOFA Score
No hypotension				0
Mean Arterial Pressure <70mmHg				1
Dopamine ≤5 or any dose Dobutamine				2
Dopamine ≥5 or Epineprine ≤0.1 or Norepinephrine ≤0.1				3
Dopamine ≥15 or Epineprine >0.1 or Norepinephrine >0.1				4

Note: vasopressor drug doses are in μg/kg/min.

Source: reference [16].

References

1. Bernal W, Auzinger G, Dhawan A, Wendon J. Acute liver failure. *Lancet* 2010; **376**(9736): 190-201.

2. Larson AM, Polson J, Fontana RJ, et al. Acetaminophen-induced acute liver failure: results of a United States multicenter, prospective study. *Hepatology* 2005; **42**(6): 1364-72.

3. Bernal W, Hyyrylainen A, Gera A, et al. Lessons from look-back in acute liver failure? A single centre experience of 3300 patients. *Journal of Hepatology* 2013; **59**(1): 74-80.

4. Ichai P, Samuel D. Etiology and prognosis of fulminant hepatitis in adults. *Liver Transplantation* 2008; **14 Suppl 2**: S67-79.

5. Ichai P, Legeai C, Francoz C, et al. Patients with acute liver failure listed for superurgent liver transplantation in France: reevaluation of the Clichy-Villejuif criteria. *Liver Transplantation* 2015; **21**(4): 512-23.

6. O'Grady JG, Alexander GJ, Hayllar KM, Williams R. Early indicators of prognosis in fulminant hepatic failure. *Gastroenterology* 1989; **97**(2): 439-45.

7. Bailey B, Amre DK, Gaudreault P. Fulminant hepatic failure secondary to acetaminophen poisoning: a systematic review and meta-analysis of prognostic criteria determining the need for liver transplantation. *Critical Care Medicine* 2003; **31**(1): 299-305.

8. Craig DGN, Ford AC, Hayes PC, Simpson KJ. Systematic review: prognostic tests of paracetamol-induced acute liver failure. *Alimentary Pharmacology & Therapeutics* 2010; **31**(10): 1064-76.

9. Bernal W, Donaldson N, Wyncoll D, Wendon J. Blood lactate as an early predictor of outcome in paracetamol-induced acute liver failure: a cohort study. *Lancet* 2002; **359**(9306): 558-63.

10. Esterline RL, Ray SD, Ji S. Reversible and irreversible inhibition of hepatic mitochondrial respiration by acetaminophen and its toxic metabolite, N-acetyl-p-benzoquinoneimine (NAPQI). *Biochemical Pharmacology* 1989; **38**(14): 2387-90.

11. Zein JG, Wallace DJ, Kinasewitz G, Toubia N, Kakoulas C. Early anion gap metabolic acidosis in acetaminophen overdose. *American Journal of Emergency Medicine* 2010; **28**(7): 798-802.

12. Craig DGN, Bates CM, Davidson JS, Martin KG, Hayes PC, Simpson KJ. Overdose pattern and outcome in paracetamol-induced acute severe hepatotoxicity. *British Journal of Clinical Pharmacology* 2011; **71**(2): 273-82.

13. Kumar R, Shalimar, Sharma H, et al. Persistent Hyperammonemia Is Associated With Complications and Poor Outcomes in Patients With Acute Liver Failure *Clinical Gastroenterology and Hepatology* 2012; **10**(8): 925-31.

14. Reddy KR, Ellerbe C, Schilsky M, et al. Determinants of outcome among patients with acute liver failure listed for liver transplantation in the United States. *Liver Transplantation* 2016; **22**(4): 505-15.

15. Roberts LN, Bernal W. Management of Bleeding and Thrombosis in Critically Ill Patients with Liver Disease. *Seminars in Thrombosis & Hemostasis* 2015; **41**(5): 520-6.

16. Vincent JL, Moreno R, Takala J, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Medicine* 1996; **22**(7): 707-10.

17. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. *Lancet* 1974; **2**(7872): 81-4.

18. Bernal W, Hall C, Karvellas CJ, Auzinger G, Sizer E, Wendon J. Arterial ammonia and clinical risk factors for encephalopathy and intracranial hypertension in acute liver failure. *Hepatology* 2007; **46**(6): 1844-52.

19. Perkins W, Tygert M, Ward E. Chi-square and classical exact tests often wildly misreport significance; the remedy lies in computers. *arxiv*, 2011. (accessed.

20. Harrell FE, Jr. Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer New York; 2001.

21. O'Grady J. Timing and benefit of liver transplantation in acute liver failure. *Journal of Hepatology* 2014; **60**(3): 663-70.

22. Wong DT, Barrow PM, Gomez M, McGuire GP. A comparison of the Acute Physiology and Chronic Health Evaluation (APACHE) II score and the Trauma-Injury Severity Score (TRISS) for outcome assessment in intensive care unit trauma patients. *Critical Care Medicine* 1996; **24**(10): 1642-8.

23. Keegan MT, Gajic O, Afessa B. Severity of illness scoring systems in the intensive care unit. *Critical Care Medicine* 2011; **39**(1): 163-9.

24. Lemeshow S, Klar J, Teres D. Outcome prediction for individual intensive care patients: useful, misused, or abused? *Intensive Care Medicine* 1995; **21**(9): 770-6.

25. Barber K, Madden S, Allen J, et al. Elective liver transplant list mortality: development of a United Kingdom end-stage liver disease score. *Transplantation* 2011; **92**(4): 469-76.

26. Knaus WA, Harrell FE, Jr., Lynn J, et al. The SUPPORT prognostic model. Objective estimates of survival for seriously ill hospitalized adults. Study to understand prognoses and preferences for outcomes and risks of treatments. *Annals of Internal Medicine* 1995; **122**(3): 191-203.

27. Sniederman A, D'Agostino R, Pencina M. The Role of Physicians in the Era of Predictive Analytics. *journal of the American Medical Association* 2015; **314**: 25-6.

28. McPhail MJ, Shawcross DL, Abeles RD, et al. Increased Survival for Patients With Cirrhosis and Organ Failure in Liver Intensive Care and Validation of the Chronic Liver Failure-Sequential Organ Failure Scoring System. *Clinical Gastroenterology & Hepatology* 2015; **13**(7): 1353-60.e8.

29. Ferreira F, Bota D, Bross A, Melot C, Vincent J. Serial Evaluation of the SOFA score to predict outcome in critically ill patients. *Journal of the American Medical Association* 2001; **286**(14): 1754-58.

30. Goettler CE, Waibel BH, Goodwin J, et al. Trauma intensive care unit survival: how good is an educated guess? *Journal of Trauma-Injury Infection & Critical Care* 2010; **68**(6): 1279-87; discussion 87-8.

31. Mathurin P, Moreno C, Samuel D, et al. Early liver transplantation for severe alcoholic hepatitis. *New England Journal of Medicine* 2011; **365**(19): 1790-800.

32. Germani G, Theocharidou E, Adam R, et al. Liver transplantation for acute liver failure in Europe: outcomes over 20 years from the ELTR database. *Journal of Hepatology* 2012; **57**(2): 288-96.